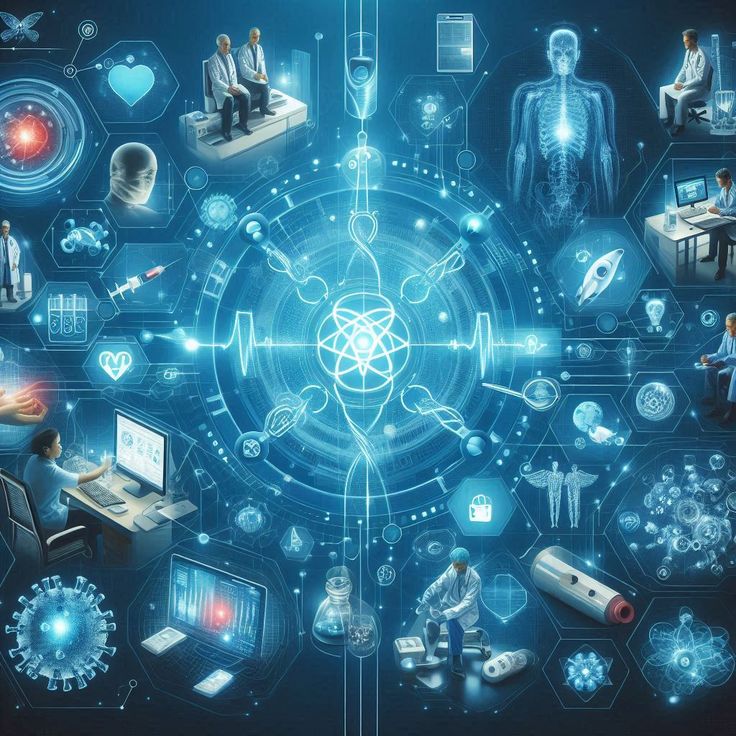
A logo of a globe and a graduation cap

AI-generated content may be incorrect.A black text on a white background

AI-generated content may be incorrect.

IBM Data Science Project

Healthcare Predictive Analytics Project



|  |  |
| --- | --- |
| Names | Eman Abdelfatah  Hazem Ibrahim  Hussein Mohamed  Kareem fekry  Mohamed Mostafa |
| Group Code | **CLS GIZ2\_AIS4\_S1** |
| Technical Instructor | **Eng. Eslam Adel** |

Table of Contents

[Project Proposal 3](#_Toc193142553)

[Problem Statement: - 3](#_Toc193142554)

[Objectives: - 3](#_Toc193142555)

[Scope: - 3](#_Toc193142556)

[Project Gantt Chart: 4](#_Toc193142557)

[Task Assignments & Roles: 4](#_Toc193142558)

[Literature Review: - 5](#_Toc193142559)

[Dataset Exploration (EDA) 6](#_Toc193142560)

[Introduction: - 6](#_Toc193142561)

[System’s Behavior: - 6](#_Toc193142562)

[Dataset Overview: - 6](#_Toc193142563)

[Data Preprocessing: - 6](#_Toc193142564)

[Exploratory Data Analysis (EDA): - 7](#_Toc193142565)

[Key Findings: - 7](#_Toc193142566)

[Machine Learning Pipeline for Health Prediction System 8](#_Toc193142567)

[Data Ingestion: - 8](#_Toc193142568)

[Data Preprocessing: - 8](#_Toc193142569)

[Feature Selection: - 8](#_Toc193142570)

[Model Training: - 8](#_Toc193142571)

[Prediction: - 9](#_Toc193142572)

[Visualization: - 9](#_Toc193142573)

[Software Architecture for Health Prediction System 10](#_Toc193142574)

[Backend: - 10](#_Toc193142575)

[Frontend: - 10](#_Toc193142576)

[Machine Learning Model: - 10](#_Toc193142577)

[Database: - 11](#_Toc193142578)

# Project Proposal

## Problem Statement: -

In the healthcare industry, timely and accurate predictions of patient health outcomes are essential for improving patient care, optimizing resource management, and supporting clinical decision-making. However, the complexity and volume of healthcare data make it challenging to derive actionable insights through traditional methods. Predictive analytics, powered by machine learning, offers a potential solution by enabling the analysis of large datasets to identify patterns and forecast patient outcomes.

This project aims to develop a predictive analytics system that can effectively process healthcare data to predict patient risks, identify trends, and support decision-making processes. This system aims to enhance patient care by providing accurate predictions that aid in early diagnosis, treatment planning, and resource allocation.

## Objectives: -

1. **Accurate Patient Risk Prediction:** Develop models that can predict patient risks based on various health metrics such as age, medical history, test results, and treatments.
2. **Trend Identification:** Analyze healthcare data to discover trends and patterns that can inform preventive measures and personalized treatment plans.
3. **Decision Support System:** Provide healthcare professionals with insights that assist in making informed decisions about patient care.
4. **Resource Optimization:** Improve the allocation of medical resources by identifying high-risk patients who may require immediate attention.
5. **Scalability & Deployment:** Build a scalable architecture that allows easy integration of new data sources and models.

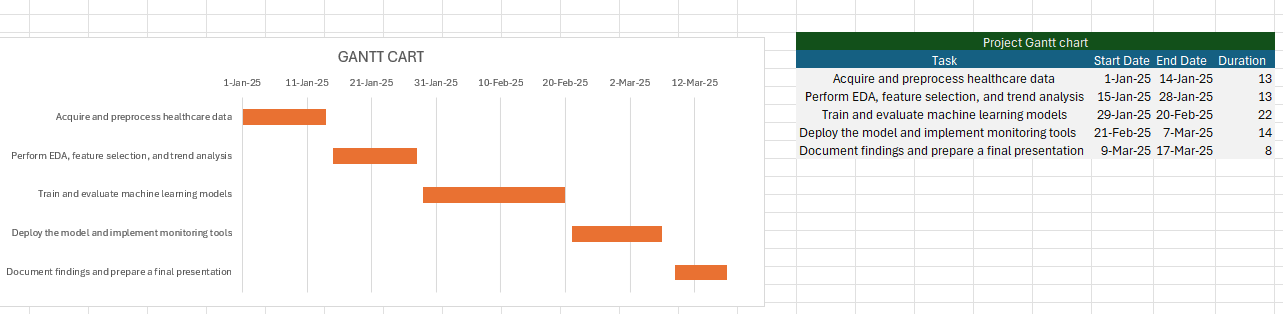
## Scope: -

1. **Data Collection**: Obtaining structured healthcare data, including patient demographics, medical history, test results, and treatment records.
2. **Exploratory Data Analysis (EDA)**: Identifying trends, correlations, and key features influencing healthcare outcomes.
3. **Model Development**: Implementing and optimizing machine learning models for accurate predictions.
4. **Deployment**: Deploying the predictive model as an interactive web application or API.
5. **Monitoring & Maintenance**: Ensuring model accuracy over time and retraining with updated data.

## Project Gantt Chart:

A graph on a sheet of paper

AI-generated content may be incorrect.



## Task Assignments & Roles:

|  |  |
| --- | --- |
| Team Member | Responsibilities |
| Mohamed Mostafa | Data collection, preprocessing, and feature engineering |
| Hazem Ibrahem | Exploratory data analysis, visualization, and model training |
| Hussein Mohamed | Model development, optimization, and deployment |
| Kareem Fekry | API development, web dashboard, and integration |
| Eman Abdelfatah | Coordination, documentation, and final reporting |

# Literature Review: -

# Dataset Exploration (EDA)

## Introduction: -

This report provides a comprehensive analysis of the Sleep Health and Lifestyle dataset. The purpose of this analysis is to explore the dataset, identify key patterns, relationships, and potential issues, and provide insights that can be used for further analysis or predictive modeling.

## System’s Behavior: -

The system follows a structured flow:

1. **Data Collection:** Data is gathered from various sources (e.g., questionnaires, surveys, sensors).
2. **Data Storage:** The collected data is stored in a structured format, such as CSV files or a database.
3. **Data Preprocessing:** Data cleaning, encoding, normalization, and handling of missing values are performed.
4. **Feature Selection:** Relevant features are selected to enhance model performance.
5. **Model Training:** Machine learning models are trained using the processed data.
6. **Prediction:** The trained models are used to make predictions based on new inputs.
7. **Visualization:** Results and insights are visualized for better interpretation.

## Dataset Overview: -

The dataset contains information related to individuals’ sleep habits, health status, and lifestyle choices. The primary features include:

* Demographic Information (e.g., Age, Gender)
* Health Indicators
* Sleep Patterns
* Lifestyle Factors

## Data Preprocessing: -

Several steps were taken to clean and prepare the data for analysis:

* Handling missing values by imputing or removing them as appropriate.
* Encoding categorical variables into numerical formats for analysis.
* Standardizing numerical features where necessary.
* Removing irrelevant columns or features with low variance.

## Exploratory Data Analysis (EDA): -

The EDA process involved generating various visualizations to better understand the data:

* **Distribution Plots:** To visualize the distribution of numerical features.
* **Count Plots:** To analyze categorical feature frequencies.
* **Pair Plots:** To explore relationships between features.
* **Correlation Analysis:** Identifying highly correlated features.

## Key Findings: -

* Some features show a significant correlation with sleep quality.
* Lifestyle factors such as smoking and alcohol consumption appear to affect sleep duration and quality.
* Stress levels are strongly associated with reduced sleep quality.
* Certain demographic factors, such as age, influence sleep patterns.

# Machine Learning Pipeline for Health Prediction System

The goal of this pipeline is to create a machine-learning model capable of predicting health conditions based on sleep patterns, lifestyle choices, and other relevant factors. The steps are structured to ensure data is properly prepared, models are trained effectively, and predictions are accurately generated.

## Data Ingestion: -

* The dataset is collected from a structured CSV file containing various features related to sleep habits, lifestyle, demographics, and health conditions.
* The data is loaded using Python libraries like Pandas for efficient manipulation.

## Data Preprocessing: -

* Handling Missing Values: Missing entries are detected and either removed or filled using appropriate techniques (e.g., mean/mode imputation).
* Encoding Categorical Data: Non-numeric columns are converted into numerical formats using encoding techniques like One-Hot Encoding or Label Encoding.
* Feature Scaling: Continuous numerical features are scaled to ensure consistent ranges for improved model performance.
* Data Splitting: The dataset is divided into training and testing subsets to evaluate model performance accurately.

## Feature Selection: -

* Important features related to predicting health conditions are selected through techniques such as correlation analysis, feature importance ranking, or Recursive Feature Elimination (RFE).
* The goal is to eliminate irrelevant or redundant features to enhance model efficiency.

## Model Training: -

* Several machine learning models are trained, including Logistic Regression, Decision Trees, Random Forest, Support Vector Machines, and Neural Networks.
* Hyperparameter tuning is performed to optimize model performance.
* Cross-validation is applied to prevent overfitting and improve generalization.

## Prediction: -

* The trained models are used to predict health conditions based on new, unseen data.
* Evaluation metrics such as Accuracy, Precision, Recall, F1-Score, and ROC-AUC are used to assess model performance.

## Visualization: -

* Results are visualized using graphs and charts to highlight important features, model performance, and predictions.
* Visualization techniques include heatmaps, bar charts, confusion matrices, and ROC curves.

# Software Architecture for Health Prediction System

The architecture of the Health Prediction System involves four main components: Backend, Frontend, Machine Learning Model, and Database. The system will be developed using Python with Streamlit for deployment, providing a user-friendly interface.

## Backend: -

* **Purpose:** Handles data preprocessing, feature engineering, model training, and prediction requests.
* **Technologies Used:** Python (Pandas, Scikit-Learn, TensorFlow/Keras, etc.).
* **Tasks Performed:**
  + Data Ingestion: Loading datasets.
  + Data Preprocessing: Handling missing values, encoding, scaling.
  + Model Training: Building and training models.
  + Prediction API: Exposing model predictions via Streamlit backend.

## Frontend: -

* **Purpose:** Provides a user interface for interacting with the prediction system.
* **Technologies Used:** Streamlit (Python Framework).
* **Tasks Performed:**
  + Data Input: Allowing users to upload new datasets or input data manually.
  + Visualization: Displaying model performance metrics and prediction results.
  + User Interaction: Providing feedback and explanation of results.

## Machine Learning Model: -

* **Purpose:** The core component responsible for making predictions.
* **Technologies Used:** Scikit-Learn, TensorFlow/Keras (if Deep Learning is applied).
* **Tasks Performed:**
  + Model Building: Selecting appropriate algorithms (Logistic Regression, Decision Trees, etc.).
  + Model Training: Training with preprocessed data.
  + Model Evaluation: Assessing performance using metrics like Accuracy, Precision, Recall, F1-score.

## Database: -

* **Purpose:** Stores datasets, trained models, and prediction results.
* **Technologies Used:** CSV files for data storage, with potential use of SQLite or other databases for scalability.
* **Tasks Performed:**
  + Data Storage: Keeping raw and processed datasets.
  + Model Storage: Saving trained models for future predictions.
  + Result Storage: Recording prediction outcomes for analysis.

The system will be deployed using **Streamlit**, ensuring simplicity and accessibility for end-users without needing complex installations.